Трубачев О. Н. Опыт ЭССЯ: к 30-летию с начала публикации (1974–2003) : доклад пленарного заседания XIII Междунар. съезда славистов в Любляне. М., 2003.

ЭССЯ — Этимологический словарь славянских языков: праславянский лексический фонд. М., 1974–. Вып. 1–.

Ю. А. Шкураток, Ю. Р. Айдаров

Пермский государственный гуманитарно-педагогический университет, Пермь
shkuratok@mail.ru, yuriy.aydarov@gmail.com

Электронный архив мифологических рассказов Пермского края: реализация системы интеллектуального поиска*

Этнолингвистические исследования, в отличие от работ, проводимых на основе книжного языка, испытывают значительные трудности на этапе поиска материала. Изучение традиционной народной культуры, отраженной в языке, требует организации полевых выездов и обработки их результатов; данные собираются по крупицам из диалектных словарей и картотек, фольклорных и этнографических записей, различного рода архивов.

Интерес к культуре и быту крестьян, активная собирательская работа, проводившаяся в течение полутора веков, привели к накоплению большого количества полевых материалов. В Пермском государственном университете за годы существования обязательной фольклорной практики была собрана коллекция песенного фольклора, быличек и других фольклорных жанров. Опубликована лишь небольшая часть текстов, остальные хранятся в виде экспедиционных тетрадей на кафедрах, в личных архивах и т. п.

Несмотря на то, что в последние годы издательская деятельность в этом направлении активизировалась, и сейчас широкому кругу исследователей доступно лишь немногое. Стоит также отметить, что бумажные сборники имеют существенный недостаток: используя их, затруднительно решать задачи, связанные с выборкой нужной

* Работа выполнена в рамках проекта РГНФ № 14-34-01279а2.

информации из массивов данных. Таким образом, можно констатировать не только проблему доступности материалов, но и проблему создания электронных архивов — вспомогательных исследовательских инструментов, облегчающих работу ученого.

К формированию фольклорных баз данных, содержащих десятки тысяч текстов, привлекаются команды компьютерных лингвистов. Результатом этого сотрудничества является оснащение баз текстов сложным поисковым аппаратом, системами визуального представления и пр. Создание больших фольклорных архивов имеет принципиальную сложность, связанную с природой таких текстов — они не имеют названия и автора в привычном понимании этого слова. Поэтому сложнейшей задачей является проблема разработки системы поиска.

Одно из решений этой проблемы связано с представлением текста в виде тезауруса ключевых слов. Этот подход имеет свои недостатки: отбор ключевых слов осуществляется вручную и не лишен субъективности. Перспективным видится автоматизация отбора ключевых слов, так как зачастую ключевые слова, отобранные вручную, уже содержатся в самих текстах [см.: Trieschnigg и др., 2013].

Особое место в создании фольклорных архивов занимает разработка инструментария, ориентированного на решение задач классификации и категоризации. Наиболее эффективными на практике считаются следующие пять подходов: частотный анализ, иерархическая онтология, использование существующей системы индексирования текстов, «топонимический» подход, персональный подход [см.: Abello и др., 2012].

Частотный анализ предполагает подсчет количества определенных элементов словаря. Его недостатком в контексте изучения фольклорных текстов на русском языке является наличие диалектных вариантов, значительно усложняющих задачу автоматизированной обработки.

Использование иерархических онтологий предполагает использование многоуровневой системы категорий, основанной на одной из разработанных в фольклористике структурных схем описания текста. Обработка текстов в рамках этого подхода выполняется вручную, что значительно затрудняет развитие метода.

В том случае, если коллекция текстов создавалась в рамках научной школы или под руководством конкретного исследователя, она может содержать определенную систему индексирования, которую возможно

использовать в системе поиска. В рамках «топонимического» подхода учитываются локации, упоминающиеся в тексте, а также населенный пункт, в котором текст был записан. Персональный подход основывается на учете имен исполнителей, а также других собственных имен, упомянутых в самих текстах.

Подводя итог, можно сделать следующие выводы.

1. Богатые фольклорные материалы, собранные на протяжении десятилетий экспедиционной деятельности, ожидают введения в научный оборот. В Пермском крае собраны тысячи текстов мифологических рассказов — чрезвычайно ценный материал ввиду хорошей сохранности пермской традиции.

2. На современном этапе развития технологий видится рациональным создание электронного архива.

3. Создание такого архива требует разработки вспомогательных исследовательских инструментов. Применение к текстам мифологических рассказов ряда апробированных в компьютерной лингвистике методик позволит создать систему интеллектуального поиска и визуального представления текстов. Это будет способствовать как введению в научный оборот неопубликованных материалов, так и решению исследовательских задач, связанных с выявлением закономерностей в текстовых массивах.

Abello J., *Broadwell P.*, *Tangherlini T. R.* Computational folkloristics // Communications of the ACM. 2012. 55 (7). P. 60–70.

Trieschnigg D., *Nguyen D.*, *Theune M.* Learning to Extract Folktale Keywords // Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities at ACL. 2013. P. 65–73.